Lang Resources & Evaluation (2015) 49:857–882
DOI 10.1007/s10579-015-9299-2

ORIGINAL PAPER

# Multimodal corpus of multiparty conversations in L1 and L2 languages and findings obtained from it

**Seiichi Yamamoto[1] · Keiko Taguchi[1] · Koki Ijuin[1] · Ichiro Umata[2] · Masafumi Nishida[1]**

Published online: 19 March 2015


**Abstract** To investigate the differences in communicative activities by the same interlocutors in Japanese (their L1) and in English (their L2), an 8-h multimodal corpus of multiparty conversations was collected. Three subjects participated in each conversational group, and they had conversations on free-flowing and goal-oriented topics in Japanese and in English. Their utterances, eye gazes, and gestures were recorded with microphones, eye trackers, and video cameras. The utterances and eye gazes were manually annotated. Their utterances were transcribed, and the transcriptions of each participant were aligned with those of the others along the time axis. Quantitative analyses were made to compare the communicative activities caused by the differences in conversational languages, the conversation types, and the levels of language expertise in L2. The results reveal different utterance characteristics and gaze patterns that reflect the differences in difficulty felt by the participants in each conversational condition. Both total and average durations of utterances were shorter in their L2 than in their L1 conversations. Differences in eye gazes were mainly found in those toward the information senders: Speakers were gazed at more in their second-language than in their native-language conversations. Our findings on the characteristics of conversations in the second language suggest possible directions for future research in psychology, cognitive science, and human–computer interaction technologies.

**Keywords** Multiparty conversation in L2 · Proficiency · Eye gaze · Multimodal corpus · Annotation

✉ Seiichi Yamamoto
seyamamo@mail.doshisha.ac.jp

[1] Faculty of Science and Engineering, Doshisha University, 1-3 Miyakodani, Tatara, Kyotanabe-shi, Kyoto 610-0321, Japan

[2] National Institute of Information and Communications Technology, 3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan

## 1 Introduction

In typical human–human interactions, the interlocutors use not only speech and language but also a wide variety of paralinguistic means and nonverbal behaviors to signal their speaking intentions to the partner, to express intimacy, and to coordinate their conversation (Argyle et al. 1968; Beattie 1978, 1980; Clark 1996; Kendon 1967; Kleinke 1986; Mehrabian and Wiener 1967; Mehrabian and Ferris 1967). Recently, there has been growing interest in the automatic analysis of conversational data, particularly the automatic analysis of multiparty conversations. Related studies have been launched within several communities, including those of human–computer interaction, machine learning, speech processing, and computer vision, with the aim of furthering our understanding of human–human communication and multimodal signaling of social interactions (Gatica-Perez 2009; Pentland 2005; Vinciarelli et al. 2009). In such multiparty conversations as a group of people informally chatting with each other or people attending a more formal meeting, it is obvious that the coordination and interaction cannot be managed in a similar way as how it is done in dialogues between two speakers who share the responsibility for coordination. Various multimodal corpora in multiparty conversations have been collected as fundamental research resources for developing automatic analysis technologies (Carletta et al. 2005; Garofolo et al. 2004).

Thanks to advanced technology, it is possible to study communicative behavior and social signaling patterns in multiparty conversations using automatic analysis techniques on these multimodal corpora. Speaker diarization (the process of partitioning an input audio stream into homogeneous segments according to speaker identity), when used together with speaker recognition (the identification of speakers by their voices), has become an important key technology for tasks such as navigation, retrieval, and high-level inference from audio data in meeting recordings. Some speaker diarization systems integrate motion and gazing data analyses with audio data analysis to achieve higher accuracy and robustness (Anguera et al. 2012; Moattar and Homayounpour 2012). There are also meeting systems that use multimodal data including both motion and gaze (Hain et al. 2010; Tur et al. 2008). Besides speech recognition, motion capture and gesture recognition technology can be used for automatic analysis of multimodal data, and the developments in eye-tracker technology allow us to study gaze behavior in an objective manner. Many quantitative studies on human–human interaction have also reported that eye gaze plays an important role in monitoring conversation content and contributes to the performance of collaborative tasks requiring the understanding of communication partners (Boyle et al. 1994; Clark and Krych 2004; Jokinen et al. 2013). These findings on human–human interactions were mainly obtained from conversations held in the mother tongue (L1). Foreign languages are used by people who travel around the world for business or pleasure and when they chat through the Internet with those living in other countries. Second-language (L2) conversations are commonly observed in daily life, and the proficiency of conversational participants typically ranges from low to high. Such differences in the proficiency of participants can cause serious miscommunications and may disrupt collaboration by both native and non-native speakers in human–human communication (Beyene et al. 2009). Uneven proficiency in

L2 may also lead to uneven opportunities for participation in conversations. A multiparty conversation consists of "ratified participants" (Goffman 1976), and participants with poorer proficiencies might be relegated to "side participant" status regardless of their level of expertise in the tasks they are working on collaboratively. It is therefore an urgent issue to develop technologies for monitoring the understanding and the contributions of all participants and for supporting smooth interactions in L2 conversations. To achieve this aim, we need to extend automatic analysis techniques to human–human interactions where participants are conversing in L2.

However, few multimodal corpora of conversations involving L2 usage have been constructed for analyzing communicative behavior and social signaling patterns, which are assumed to be different from those in strictly L1 conversations. Furthermore, there has been a near-total lack of multimodal corpora made from L1 + L2 conversation by the same interlocutors to precisely analyze the differences in their L1 and L2 communicative behaviors.

Our previous studies in multiparty conversations suggested that gazing activities, one of the most important features for monitoring conversation content and understanding communication partners, were different between conversations in L1 and L2: the gazing duration by listeners in conversations in L2 are longer than those in L1, although there was no such difference in the gazing activity of the speaker (Yamasaki et al. 2012; Kabashima et al. 2012). Although these previous studies suggested differences in communicative behaviors, such as utterances and eye gazes, between the conversations in L2 and those in L1, the data size was insufficient to take into consideration the various factors expected to affect communicative behaviors, such as L2 expertise and conversational topic.

To investigate the differences in communicative activities by the same interlocutors in Japanese (their L1) and in English (their L2), an 8-hour multimodal corpus of multiparty conversations was collected. Three subjects at various levels of conversational expertise in L2 participated in each conversational group, and they had conversations on free-flowing and goal-oriented topics in L1 and L2. Their utterances, eye gazes, and gestures were recorded with three microphones, eye trackers, and video cameras. We collected a total of 80 conversations by 20 conversational groups. Quantitative analyses were conducted to compare differences in utterances and in eye gaze activities caused by the differences in conversational languages, the conversation types, and the participants' levels of L2 language expertise.

This paper is structured as follows. We describe the multimodal data collected in this research in Sect. 2 and continue with annotation and transcription of the data for corpus creation in Sect. 3. We report our analytical results on eye gaze and the characteristics of utterances in Sect. 4 and present a discussion in Sect. 5. Our conclusions are given and future works are discussed in Sect. 6.

## 2 Data collection

It has been shown in previous research that mutual gaze is important in the coordination of interaction (e.g., Kendon 1967; Argyle and Cook 1976), but these studies mainly dealt with two-party dialogues, not multiparty conversations. We

collected multimodal data in three-party conversation to investigate whether the role of eye gaze is as important in three-party conversation as in two-party dialogue (Jokinen et al. 2013). Furthermore, we collected 80 multimodal datasets (20 free-flowing in Japanese, 20 free-flowing in English, 20 goal-oriented in Japanese, and 20 goal-oriented in English) from multiparty conversations to investigate the difference in eye gaze between L1 and L2 conversations. The same twenty groups of three participants each conversed in all four conditions. The average duration of a dataset was 6 min. Table 1 lists quantitative features of the collected data.

## 2.1 Experimental setup

Three subjects participated in each conversational group and sat in a triangular formation around a table. The distance between the three participants was 1.5 m (Fig. 1). Three sets of eye trackers, headsets with microphones, and video cameras were used, and the eye gazes, voices and gestures of all three participants were recorded (Fig. 2). A start signal triggered these instruments in synchronization with each other.

**Table 1** Quantitative features of collected data

| Features | Numerical values |
|---|---|
| Total participants | 20 conversational groups of 3 participants |
| Average duration of conversations | 6 min |
| Conversational types | 2 types (free-flowing, goal-oriented) |
| Conversational languages | Japanese (L1), English (L2) |
| TOEIC scores of participants | 450–985 points |
| Annotated values | *Speech* (start and end times), *GazeObject*, *Turn* |

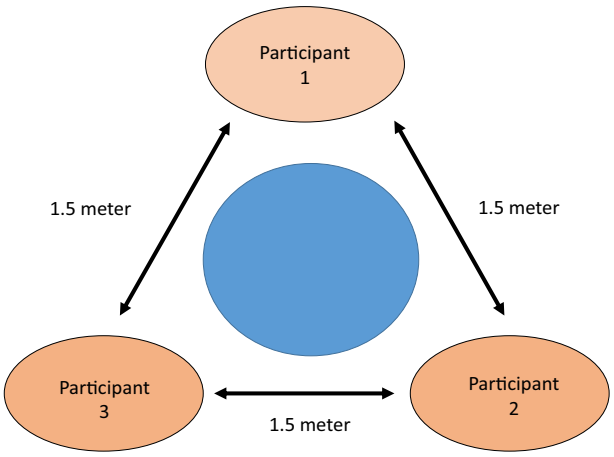


**Fig. 1** Seating positions for three participants during collection of multiparty conversation

**Fig. 2** Experimental setup

Students of Doshisha University were recruited to participate in the experiment, and all sessions were conducted in the same lecture room at this university.

We used NAC EMR-9 (NAC Image Technology Inc.) eye trackers in this experiment. These were cap-mounted eye-tracking systems that enabled the participants to move their heads and hands freely in accordance with their conversational activities. The eye-tracker had one eyesight camera and two eye cameras as well as near-infrared sensors. The eyesight camera recorded the scene that the subject was gazing at. The angle of view was 62°. The eye cameras recorded the eye movements of all participants at a sampling rate of 60 fps. The near-infrared sensors recorded the participants' pupils and a figure that was reflected from the cornea. Their eye gazes were not tracked when they blinked, when they laughed, or when their eyes had narrowed so much that the eye tracker could not detect their pupils.

## 2.2 Participants

A total of 60 subjects (20 groups) between the ages of 18 and 24 participated in this experiment as previously explained. They were Japanese university students who had acquired Japanese as their L1 and had learned English as their L2. They were not acquainted with each other before the meeting held for data collection. Their communication levels in English were measured using the Test of English for International Communication (TOEIC). Their scores ranged from 450 to 985 (990 being the highest score that could be attained). Figure 3 denotes the cumulative distribution of their TOEIC scores in comparison with that in the latest TOEIC test administered nationwide (TOEIC_TEST). Both cumulative distributions show nearly the same figure, although the cumulative distribution values of TOEIC_DATA are slightly higher in the lower TOEIC ranges than those of the participants.

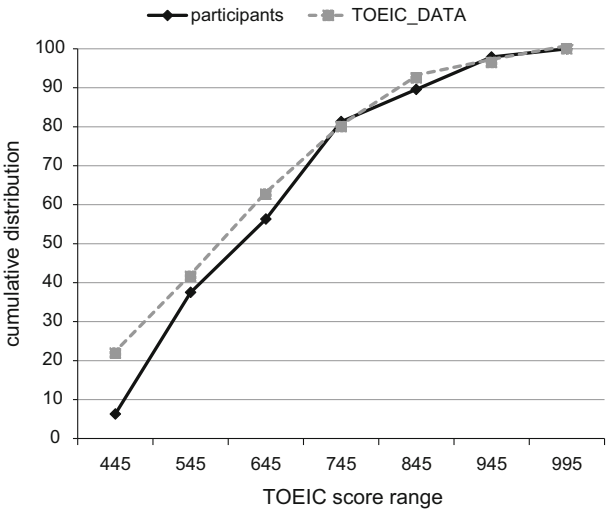


**Fig. 3** Cumulative distribution of participants' TOEIC scores

The differences in eye gaze and utterance between L1 and L2 conversations may depend on their L2 proficiencies themselves or on the relative difference among their L2 proficiencies in the conversational group. Therefore, we recruited participants of various L2 proficiencies to make up the conversational groups. Accordingly, we assembled groups of various combinations of L2 proficiencies, such as groups of participants with high TOEIC scores, those with low scores, and those with high/middle/low scores.

## 2.3 Procedure

In the procedure used for the experiment, the content of the conversational topic was first explained to the participants. There were two conversation types. The first was free-flowing, natural chatting that covered various topics such as hobbies, weekend plans, studies, and travel. The second type was goal-oriented, in which participants collaboratively decided what to take with them on trips to uninhabited islands or mountains. We randomly arranged the order of the conversation types to cancel out the effect of order. We also randomly arranged the order of the languages used in the conversations. After the experiment had been explained, eye trackers were calibrated. During calibration, the participants looked at nine points on the calibration board while the system measured the eye's position and shape and the reflections of the infrared light. This calibration was done for each participant in parallel, and participants started conversing after all of them finished the calibration. Each group had conversations about free-flowing and goal-oriented topics in Japanese and in English. Furthermore, the participants filled out a questionnaire after each conversation. Each question was categorized into features that categorize communication, such as participants' gazing activities, their feelings toward other

**Table 2** Annotation features and values in previous research

| Annotation features | Feature values |
|---|---|
| *DialogAct* | Backchannel, stall, fragment, bepositive, benegative, suggest-offer, inform, ask, other |
| *GazeObject* | RS (Gaze at the person to your right), LS (gaze at the person to your left), other (gaze to other), nogaze |
| *Head Movement* | Nod, jerk, backward, forward, tilt, TurnToPartner, TurnSide, waggle, other |
| *Head Repetition* | Single, repeated, none |
| *Handness* | Both, single |
| *Trajectory Right Hand* | Forward, backward, side, up, down, complex, other |
| *Trajectory Left Hand* | Forward, backward, side, up, down, complex, other |
| *HandRepetition* | Single, repeated, none |
| *Turn* | Give, take, hold |

participants, their interest in the topic of conversation, their conversational skill in English, and their evaluation of the conversation content (Umata et al. 2013). Consequently, the subjects in each group participated in four conversations and filled out four questionnaires.

## 3 Corpus creation

### 3.1 Annotation features

For this analysis, we used the EUDICO Linguistic Annotator (ELAN) developed by the Max Planck Institute for Psycholinguistics (MPIP), which is a linguistic annotation tool for creating text annotations onto video and audio files. We performed the annotation according to the MUMIN annotation scheme (Allwood et al. 2007) used in our previous research for modeling turn-taking behaviors in L1 conversations (Jokinen et al. 2013). The annotation features and values adopted in our previous research are listed in Table 2. Our preliminary test showed that the inter-coder agreement between the annotators, which was measured by Cohen's kappa coefficient, was not high for some features such as *Dialogue Act*, *Head Movement*, and *Hand Movement* in the case of conversations in L2. Cohen's kappa coefficients were 0.55, 0.14, and 0.34 for *Dialogue Act*, *Head Movement*, and *Hand Movement*, respectively, in our preliminary test. We decided to start making annotations by limiting them, in the first stage of the research, to the features that were reported to be important in monitoring conversations, such as utterances, eye gazes, and turn-taking activities (Jokinen et al. 2013). These features also maintained high agreement among the annotators. The authors manually determined the start and end times of each utterance by considering only pause durations (500 ms) between consecutive speech segments, that is, not considering the contents of consecutive speech segments. The authors adopted the annotation feature *TURN*

for turn-taking activity such as turn-give, turn-take, and turn-hold, based on the pause duration between consecutive utterances.

### 3.2 Gaze events

As already mentioned, many quantitative studies have reported that eye gaze plays an important role in monitoring conversation content and contributes to the performance of collaborative tasks needing the understanding of communication partners. The authors also selected gaze events as one of the annotation features and manually annotated the feature *GazeObject* based on the gaze path given by the eye tracker to obtain a more precise annotation feature; moreover, "Gaze at the person to your right," "Gaze at the person to your left," "Gaze to other," (Gaze to objects besides the person to your right or left) and "NoGaze" (Gaze was not detected) were used as the values of *GazeObject*. Gaze events are defined as gazing at some object, that is, the participant focuses her visual attention on a particular object for a certain period of time (more than 200 ms). In the current study, there are three named objects: the two partners and the "other." While focusing one's visual attention on something can also include gaze shifts to the whole context, we concentrate on the local attention level and thus count each gaze shift as a separate gaze event. With gaze events we have to note, however, that small movements around the same gaze object are included in one continuous gaze event. There are two reasons for this kind of eye movement: the so-called saccades, which refer to the eyes' involuntary and constant movements of their fixation points, and the agent's own involuntary eye movement around the gaze object while generally focusing on the object. We also include the shifts from one object to the other in the event, within the outline of the gaze object, that is, once the gaze is shifted beyond the outlines of the current gaze object, the gaze event is also changed to another. Gaze signals may also be broken in that there is no signal data. This is due to technical reasons but also to the agent's visual attention not changing (within 0.2 s). Furthermore, it is clear that when the agent's visual attention has not changed, the elements are considered part of the same gaze event; otherwise, they are considered two different gaze events.

We tried automatic analyses of eye tracking signals, but we decided to manually annotate the features to obtain more precise annotation because the low resolution of the eye tracker prevented reliable automatic detection of faces and bodies. The authors, therefore, manually annotated the start time and end time of gaze events and their feature values based on the previously described procedures. That is, the time when the gaze shifts within the outline of the gaze object is set to the start time of the gaze event to the object if the gaze does not shift beyond the outlines of the gaze object in less than 200 ms. The end time of the gaze event is set to the time when the gaze shifts beyond the outline of the gaze object if this time is more than 200 ms. Figure 4 shows an annotation screenshot where the video is shown at the top and the annotations are added to the rows at the bottom. Cohen's kappa coefficient of segmenting gaze events was 0.83.

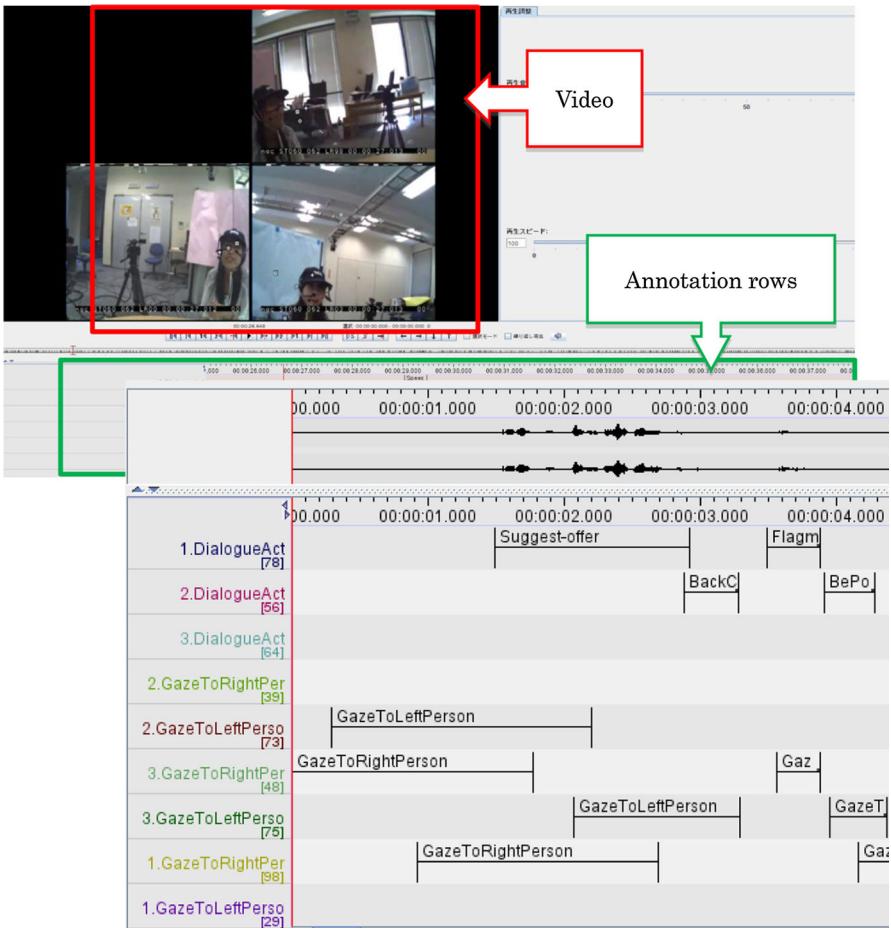


**Fig. 4** Screenshots of executing ELAN and enlargement of its *annotation rows*

### 3.3 Transcription

Limited vocabulary of non-native speakers forces speakers to express themselves in unsuitable words and non-native speech usually includes less fluent pronunciation as well as mispronunciation even in cases where it is well composed. It was difficult even for native speakers to correctly transcribe the recorded speech with these features even if they can understand speech spoken by non-native speakers in real conversational situations.

After all conversations finished, the recorded voices in the conversations in L2 were transcribed by the participants themselves and checked with a bilingual assistant. The transcription procedures were specified by the authors. For example, when the speaker was laughing or hesitating, the span had to be surrounded by an exclamation point as in !laugh!. Words also had to be bounded by hash marks as

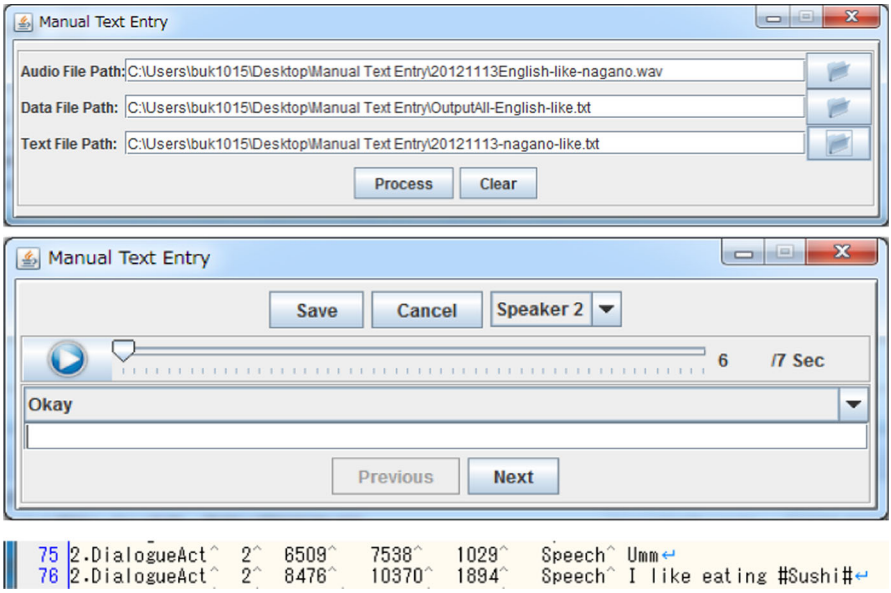


**Fig. 5** Interface for linking annotated tags of utterances and their transcribed data, with an example of linked data. *Each column* denotes line number, value of dialogue act (not yet annotated), ID of participant, start, end, and duration of each utterance, and transcription

in #Tokyo# or #sushi# when speakers uttered a proper noun or a word in Japanese. We developed a tool for linking annotated tags of utterances and their transcribed data. Figure 5 shows the interface of the software tool and examples of linking annotated tags of utterances and their transcribed speech by a participant.

The transcribed utterances of each participant were time-aligned with those of other participants along the time axis based on the linked results to analyze the relationship between the content of utterances and differences in movements of eye gazes in each utterance of various dialogue acts.

## 4 Features of conversations

In the following, we describe some features of utterances and eye gazes that were obtained using the conversational data.

### 4.1 Utterances

Some studies have regarded "nativeness" as "expertise" and compared the grounding process between differing levels of language expertise (Kasper 2004; Hosoda 2006). Based on these ideas, we expected that the linguistic expertise of the participants in L2 would be varied and, moreover, the difficulty in L2 conversation would be greater for participants with lower linguistic expertise in L2. Our

assumption was that there was little difference in their linguistic expertise in L1. We also assumed that their linguistic expertise in L2 could be measured by their TOEIC scores and labeled participants with the highest TOEIC scores Rank1, those with the second-highest TOEIC scores Rank2, and those with the third-highest TOEIC scores Rank3 in each conversational group. These rankings were used to analyze the level of difficulty owing to linguistic expertise.

We predicted that the participants would speak more in the L1 conversations. We compared (i) the percentage of silence in the conversation (ii) the number of utterances, and (iii) total utterance duration and average utterance duration between the L1 and L2 conversations as indices of participants' difficulties in communication in L2. We also analyzed the effect of conversation type and the expertise of the participants on the difficulty of communication in L2.

Table 3 lists basic statistics of the percentage of silence duration, the total utterance duration (TUD), the average utterance duration (AUD), and the number of utterances in four kinds of conversations, i.e., those on goal-oriented topics in L1 and L2 and free-flowing ones in L1 and L2. Figures 6, 7, and 8 show the total and average utterance durations and the number of utterances of the Rank1, Rank2, and Rank3 participants.

## 4.2 Analyses of utterances

The percentage of silence duration was the greatest in the conversation on the goal-oriented topics in L2, and the smallest in the free-flowing ones in L1. Under the hypothesis that the percentage of silence was greater in conversations in L2 than in L1 and that the conversation type affected the percentage of silence, we conducted an ANOVA test to compare the percentage of silence duration among each group, with both the language difference and the conversation type difference being

**Table 3** Basic statistics of utterances

| Features in conversation | Average ± SD | | | |
|---|---|---|---|---|
| | Free (JPN) | Free (ENG) | Goal (JPN) | Goal (ENG) |
| Percentage of silence duration | 25.9 ± 9.7 | 45.4 ± 11.9 | 34.7 ± 9.3 | 52.2 ± 11.6 |
| Total utterance duration (TUD) | 101.8 ± 34.6 | 68.8 ± 32.0 | 93.6 ± 40.2 | 61.2 ± 35.9 |
| TUD of Rank1 (s) | 102.0 ± 36.9 | 71.7 ± 33.9 | 96.7 ± 43.9 | 75.1 ± 43.4 |
| TUD of Rank2 (s) | 104.3 ± 28.6 | 78.2 ± 29.4 | 83.5 ± 30.0 | 68.3 ± 29.8 |
| TUD of Rank3 (s) | 98.9 ± 39.1 | 56.5 ± 30.1 | 100.6 ± 45.0 | 40.2 ± 22.7 |
| Average utterance duration (AUD) | 1.53 ± 0.45 | 1.26 ± 0.41 | 1.43 ± 0.48 | 1.14 ± 0.47 |
| AUD of Rank1 (ms) | 1.53 ± 0.37 | 1.21 ± 0.40 | 1.44 ± 0.35 | 1.22 ± 0.58 |
| AUD of Rank2 (ms) | 1.55 ± 0.50 | 1.32 ± 0.38 | 1.30 ± 0.30 | 1.21 ± 0.37 |
| AUD of Rank3 (ms) | 1.50 ± 0.50 | 1.26 ± 0.46 | 1.54 ± 0.69 | 0.99 ± 0.42 |
| Number of utterances (NU) | 69.8 ± 24.8 | 56.8 ± 29.2 | 67.7 ± 23.1 | 54.8 ± 26.7 |
| NU of Rank1 | 70.0 ± 28.5 | 63.2 ± 35.8 | 68.4 ± 24.9 | 62.0 ± 24.8 |
| NU of Rank2 | 70.8 ± 21.4 | 61.5 ± 26.7 | 65.9 ± 21.6 | 59.8 ± 28.5 |
| NU of Rank3 | 68.5 ± 25.1 | 45.7 ± 21.5 | 68.9 ± 23.7 | 42.8 ± 23.6 |

within-subject factors. The results revealed significant main effects of both language difference ($F_{(1,\ 19)} = 125.6$, $p < .01$) and conversation type difference ($F_{(1,\ 19)} = 37.8$, $p < .01$), and no interaction was observed. This analysis result shows that silence duration is longer in conversations in L2 than in L1 and is also longer in goal-oriented conversations than in free-flowing ones. The analysis result failed to show that the language effect was stronger in goal-oriented conversations

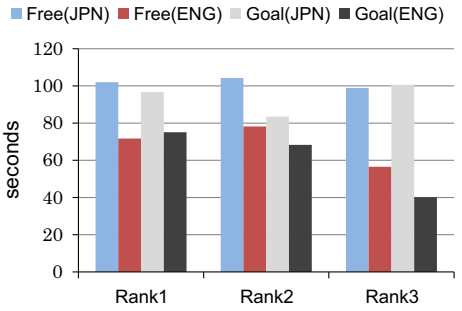


**Fig. 6** Total utterance durations of Rank1, Rank2, and Rank3 participants

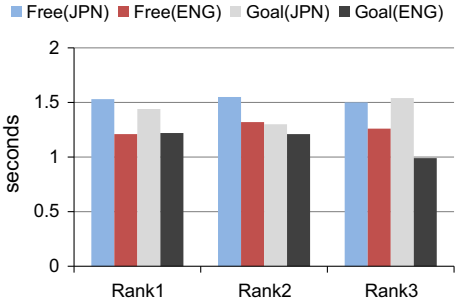


**Fig. 7** Average utterance durations of Rank1, Rank2, and Rank3 participants

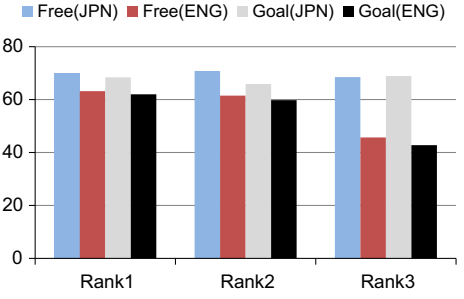


**Fig. 8** Number of utterances of Rank1, Rank2, and Rank3 participants

than in free-flowing ones, probably due to the large difference in silence duration between goal-oriented and free-flowing conversations.

We expected both the total utterance duration (TUD) and the average utterance duration (AUD) to be longer in L1 than in L2 and, moreover, the conversation type and the expertise in the L2 to affect these features of utterances. Under these hypotheses, we conducted an ANOVA test for TUD and AUD, with both the language difference and the conversation type difference being within-subject factors and L2 expertise being the between-subject factor. For TUD, the results revealed a significant main effect of language difference ($F_{(1,\ 57)} = 95.0$, $p < .01$) and a significant main effect of conversation type differences ($F_{(1,\ 57)} = 10.8$, $p < .01$). There was a significant interaction between the expertise in L2 and the language difference ($F_{(2,\ 57)} = 8.0$, $p < .01$). These analysis results show that TUD is larger in conversations in L1 than in L2 and is also larger in free-flowing conversations than in goal-oriented ones. The analysis result on interaction between expertise in L2 and the language difference shows that the decrease in TUD from conversations in L1 to those in L2 depends on the expertise of the participants and that the decrease of Rank3 was more than the others, as shown in Fig. 6.

For AUD, the results revealed a significant main effect of language difference ($F_{(1,\ 57)} = 31.5$, $p < .01$) and a significant main effect of conversation type difference ($F_{(1,\ 57)} = 6.8$, $p < .05$). This result shows that AUD is larger in conversations in L1 than in L2 and is also larger in free-flowing conversations than in goal-oriented ones. The analysis result of AUD on interaction between expertise in L2 and the language difference was different from that of TUD, and there was no significant interaction between the expertise in L2 and the language difference, although there was a second order interaction ($F_{(2,\ 57)} = 4.6$, $p < .05$) among the expertise in L2, the language difference, and conversation type difference. This result suggests that the difference among AUDs of speakers of each Rank is not so significant as that of TUD.

Under the hypotheses that the number of utterances was greater in conversations in L2 than in L1 and that the conversation type affected the number of utterances, we conducted an ANOVA test with the language difference and the conversation type difference being within-subject factors and L2 expertise being the between-subject factor. For the number of utterances, the results revealed a significant main effect of language difference ($F_{(1,\ 57)} = 37.1$, $p < .01$). There was a significant interaction between expertise in L2 and the language difference ($F_{(2,\ 57)} = 7.4$, $p < .01$). This analysis result shows that the number of the utterances is larger in conversations in L1 than in L2, but there is not so much of a difference between free-flowing conversations and goal-oriented ones. The analysis result on interaction between expertise in L2 and the language difference suggested that the decrease of TUD from conversations in L1 to those in L2 was mainly due to the decrease in the number of utterances by Rank3.

Table 4 lists the ANOVA test results of the percentage of silence durations, TUD, AUT, and the number of utterances.

**Table 4** Features of ANOVA test results on utterances

| Features | ANOVA test results | | |
|---|---|---|---|
| | Language diff. (LD) | Conversation type diff. (CD) | LD, CD, and diff. of expertise in L2 (ED) |
| Percentage of silence durations | $F_{(1,19)} = 125.6$ $P < 0.01$ | $F_{(1,19)} = 37.8$, $P < 0.01$ | n.s.i. ($F_{(2,57)} = 0.9$) |
| Total utterance duration | $F_{(1,57)} = 95.0$, $P < 0.01$ | $F_{(1,57)} = 10.8$, $P < 0.01$ | $F_{(2,57)} = 8.0$, $P < 0.01$ between LD and ED<br>$F_{(2,57)} = 5.0$, $P < 0.01$ among LD, CD, ED |
| Average utterance duration | $F_{(1,57)} = 31.5$, $P < 0.01$ | $F_{(1,57)} = 6.8$, $P < 0.05$ | $F_{(2,57)} = 4.5$, $P < 0.05$ among LD, CD, ED |
| Number of utterances | $F_{(1,57)} = 37.1$, $P < 0.01$ | n.s.m. ($F_{(1,57)} = 0.8$) | $F_{(2,57)} = 7.4$, $P < 0.01$ between LD and ED |

*n.s.i.* no significant interaction, *n.s.m.* no significant main effect

Analyses of second order interactions are listed in Appendix 1

### 4.3 Gaze events in speaking

As already mentioned, many quantitative studies on human–human interaction have reported that eye gaze plays an important role in coordinating conversations and contributes to the performance of collaborative tasks needing the understanding of communication partners. Previous research (Yamasaki et al. 2012; Kabashima et al. 2012) has suggested that eye gaze in speaking differs according to whether conversation is conducted in L1 or L2. This result suggests that the function of eye gaze in conversations in L2 might be different from that in L1. We analyzed the eye gazes of both speakers and listeners in conversations in L2. More specifically, we analyzed (1) how long the speaker was gazed at by other participants, (2) how long the speaker gazed at other participants in conversations in L2 in comparison with those in L1, and (3) whether the expertise of participants affected their gazing activities.

We used the average of gazing-at ratios to analyze how long the speaker gazed at other participants and the average of being-gazed-at ratios to analyze how long the speaker was gazed at by other participants in the previous research (Yamamoto et al. 2013). In this paper we used the speaker's gazing ratio and the listener's gazing ratio to analyze them more precisely. The speaker's gazing ratio indicates how long the speaker gazed at other participants during his/her utterances and is defined as the ratio of the duration of the speaker gazing at other participants to his/her speaking duration. The listener's gazing ratio indicates how long a participant gazed at the speaker during his/her utterance and is defined as the ratio of the duration of a participant gazing at the speaker to the speaking duration. Figure 9 illustrates the concepts of the speaker's gazing ratio and the listener's gazing ratio. The being-gazed-at ratio shows total gazing activities of both listeners, but the listener's gazing ratio indicates the gazing activities of each

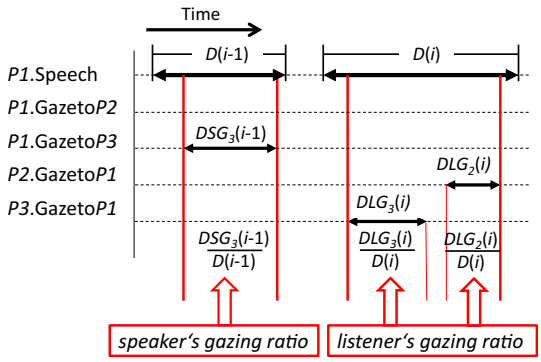


**Fig. 9** Flow diagram for calculating speaker's and listener's gazing ratios

listener independently. The speaker's gazing ratio is the same as the gazing-at ratio, but we rename the gazing-at ratio as the speaker's gazing ratio to clarify the relation between the speaker's and the listener's gazing activities. The average of the speaker's gazing ratios is defined as

$$Average\ of\ speaker's\ gazing\ ratios = \frac{1}{n}\sum_{i=1}^{n}\frac{DSG_j(i)}{D(i)} \times 100\,(\%).$$

Here, $D(i)$ is the duration of the $i$th utterance and $DSG_j(i)$ is the duration of the speaker gazing at the $j$th participants ($j = 1, 2, 3$) in the $i$th utterance.

The average of the listener's gazing ratios was defined as

$$Average\ of\ listener's\ gazing\ ratios = \frac{1}{n}\sum_{i=1}^{n}\frac{DLG_j(i)}{D(i)} \times 100\,(\%).$$

Here, $DLG_j(i)$ is the total duration of the $j$th participant ($j = 1, 2, 3$) gazing at the speaker in the $i$th utterance.

In order to evaluate whether the L2 expertise of participants affects one's gazing activities, we calculated the averages of the speaker's gazing ratios and the listener's gazing ratios, and the average listener's gazing ratios at Rank1, Rank2, and Rank3. In order to evaluate whether the differences among the L2 expertise levels of participants affects one's gazing activities, we also calculated the average gazing ratio of participants of Rank2 to Rank1, that of Rank3 to Rank1, and that of Rank3 to Rank2 as the average gazing ratio of participants of lower expertise to participants of higher expertise; likewise, we calculated the average gazing ratio of participants of Rank1 to Rank2, that of Rank1 to Rank3, and that of Rank2 to Rank3 as the average gazing ratio of participants of higher expertise to participants of lower expertise.

**Table 5** Basic statistics of eye gaze activities

| Features in conversation | Average ± SD | | | |
|---|---|---|---|---|
| | Free (JPN) | Free (ENG) | Goal (JPN) | Goal (ENG) |
| Number of gaze events | 116.9 ± 36.2 | 94.4 ± 27.2 | 106.0 ± 40.1 | 79.5 ± 31.0 |
| Speaker's gazing ratios | 0.28 ± 0.13 | 0.28 ± 0.14 | 0.28 ± 0.16 | 0.28 ± 0.16 |
| Listener's gazing ratios (LGR) | 0.47 ± 0.14 | 0.58 ± 0.15 | 0.44 ± 0.16 | 0.57 ± 0.17 |
| LGR at Rank1[a] | 0.47 ± 0.13 | 0.56 ± 0.13 | 0.44 ± 0.16 | 0.54 ± 0.17 |
| LGR at Rank2 | 0.46 ± 0.17 | 0.59 ± 0.15 | 0.43 ± 0.14 | 0.59 ± 0.16 |
| LGR at Rank3 | 0.48 ± 0.11 | 0.58 ± 0.18 | 0.43 ± 0.17 | 0.58 ± 0.17 |
| LGR from HIGH to LOW | 0.48 ± 0.13 | 0.59 ± 0.17 | 0.43 ± 0.15 | 0.60 ± 0.16 |
| LGR from LOW to HIGH | 0.46 ± 0.15 | 0.57 ± 0.14 | 0.44 ± 0.16 | 0.54 ± 0.17 |

[a] LGR at Rank1 means listener's gazing ratios in the case where a listener is gazing at a speaker of Rank1

### 4.4 Analyses of gaze events in speaking

Table 5 lists basic statistics of the number of gaze events, speaker's gazing ratios, and listener's gazing ratios. The basic statistics in Table 5 show that the number of eye gaze events in four conversational conditions have almost the same tendency as the number of utterances in Table 3, that is to say, that the larger the number of utterances is, the larger the number of eye gaze activities. This result suggested a relation between utterances and eye gaze events.

As for gazing ratios, the averages of speaker's gazing ratios seemed to be almost the same in four kinds of conversations. On the other hand, the averages of listener's gazing ratio were larger in conversations in L2 than in L1, and they were also slightly larger in the free-flowing conversations than the goal-oriented ones. Figure 10a compares averages of listener's gazing ratios at a speaker of each Rank among the four kinds of conversations, i.e., the free-flowing conversations and the conversations on goal-oriented topics in both L1 and L2. Figure 10b, c compare averages of listener's gazing ratios of participants of higher expertise to participants of lower expertise, and vice versa, among the four kinds of conversations.

Under the hypotheses that the speakers were gazed at more in conversations in L2 than in L1, and that the conversation type and the L2 expertise of participants affects one's gazing activities, we conducted an ANOVA test with the language difference and the conversation type difference being within-subject factors and the expertise of speakers being the between-subject factor. The results revealed a significant main effect of language difference ($F_{(1,\ 117)} = 107.7$, $p < .01$) and a significant main effect of conversation type difference ($F_{(1,\ 117)} = 6.0$, $p < .05$). There was no significant interaction between language difference and differences in expertise of speakers in L2.

This result shows that a speaker is gazed at more by listeners in conversations in L1 than in L2, and in the goal-oriented conversations than in the free-flowing ones in both languages.

We also conducted an ANOVA test with the language difference and the conversation type difference being within-subject factors and the difference

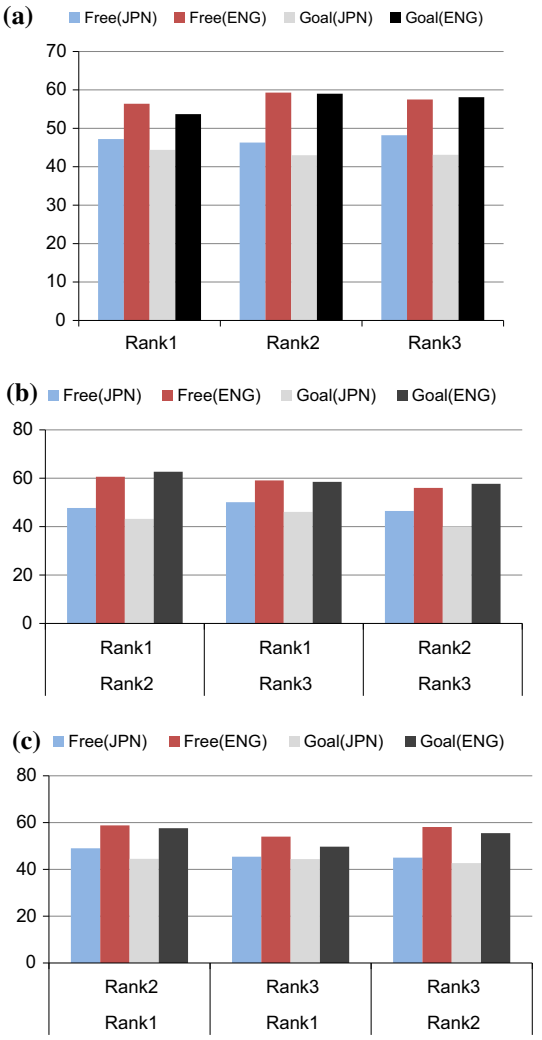


**Fig. 10** Averages of listeners' gazing ratios for various listeners in four kinds of conversations. **a** Average of listeners' gazing ratios at speaker of each Rank. **b** Average of listeners' gazing ratios of participants of higher expertise to participants of lower expertise (higher-listener vs. lower-speaker). **c** Average of listeners' gazing ratios of participants of lower expertise to participants of higher expertise (lower-listener vs. higher-speaker)

between the listener's gazing ratios of participants of higher expertise to participants of lower expertise and those of participants of lower expertise to participants of higher expertise being the between-subject factor. The results revealed a significant main effect of language difference ($F_{(1,\ 117)} = 107.4$, $p < .01$) and a significant main effect of conversation type difference ($F_{(1,\ 117)} = 6.2$, $p < .05$). There was no significant interaction between language difference and differences in expertise in L2.

As for the average of speaker's gazing ratios, we could not find any significant difference for either language difference or different conversation type.

Table 6 lists the ANOVA test results of eye gaze activities.

**Table 6** Features of ANOVA test results on eye gaze activities

| Features | ANOVA test results | | |
|---|---|---|---|
| | Language diff. (LD) | Conversation type diff. (CD) | LD, CD, and diff. of expertise in L2 (ED) |
| Average of speaker's gazing ratios | n.s.m. ($F_{(1,117)} = 0.1$) | n.s.m ($F_{(1,117)} = 0.2$) | n.s.i. ($F_{(2,117)} = 0.6$) between LD and ED |
| Average of listener's gazing ratios (LGR at Rank) | $F_{(1,117)} = 107.7$, $P < 0.01$ | $F_{(1,117)} = 6.0$, $P < 0.05$ | n.s.i. ($F_{(2,117)} = 2.1$) between LD and ED |
| Average of listener's gazing ratios (LGR from H to L/from L to H) | $F_{(1,117)} = 107.4$, $P < 0.01$ | $F_{(1,117)} = 6.2$, $P < 0.05$ | n.s.i. ($F_{(2,117)} = 1.7$) between LD and ED |

*n.s.i.* no significant interaction, *n.s.m.* no significant main effect

## 5 Discussions

We have conducted quantitative analyses on utterance and eye gaze using the multimodal corpus of multiparty conversations in L1 and L2. The main points obtained through the analyses are as follows.

### 5.1 Differences in utterances

The results in Sect. 4.2 indicated that there were significant differences of silence durations, total and average utterance durations, and the number of utterances between conversations in L1 and L2. The shorter total and average utterance durations and the longer silence in L2 suggest the difficulties the participants had in the L2 conversations. These results suggested that the participants produced shorter utterances in the conversations in L2, which are assumed to be simpler expressions; however, content analysis should be conducted to confirm this assumption.

There were significant differences in silence duration, TUD, and AUD between free-flowing conversations and goal-oriented conversations in both L1 and L2. This result suggests that the difficulty of speech production was more serious in goal-oriented conversations than in free-flowing ones, maybe because they need more time to produce speech corresponding to discourse in goal-oriented conversations. We expected that the difficulty would be more serious in conversations on goal-oriented topics in L2 than in L1 due to a shortage of vocabulary and colloquial expressions for the topic; however, we could not obtain such an analysis result. This may be due to the very strong effect of the language difference in comparison to the conversation-type difference.

As for TUD and the number of utterances there was a significant interaction between the expertise in L2 and the language difference; however, we couldn't

find a significant interaction between the expertise in L2 and the language difference for AUD. The analysis result on interaction between the expertise in L2 and the language difference shows that difference between TUD of conversations in L1 and in L2 was mainly due to the decrease in the number of utterances by Rank3.

The decreasing ratio of AUD of Rank2 from conversations in L1 to those in L2 was smaller than those of Rank1 and Rank3. We think that the phenomenon might be a kind of "alignment" (Garrod and Pickering 2004) and that participants of the middle-expertise group (Rank2) play a role of mediating conversation between participants of high and low expertise. We will conduct research on this phenomenon using the transcribed speech data and the annotation on the dialogue act used for grounding.

### 5.2 Differences in eye gazes

The results in Sect. 4.4 indicated that eye gaze in the L2 conversations were different from those in L1. The speakers were gazed at more by their listeners in the conversations in L2 than those in L1. This phenomenon was found in both free-flowing conversations and those on the goal-oriented ones in L2. Several possible reasons arise to explain the difference between gaze activities in the conversations in different languages, among them: (1) participants monitored their understanding of what was being said to make repairs if necessary, (2) participants used visual information to help in perceiving the auditory information, (3) participants gave a polite acknowledgement of the speaker's effort in producing speech with difficulty. We should investigate the cause of this phenomenon by analyzing the relation between the gaze activities and the speech act.

### 5.3 Effect of expertise in L2

The effect of different levels of L2 expertise were tested concerning features of utterances as well as eye gaze behavior. Concerning features of utterances, we found significant interactions between language difference and expertise in L2. As for the average of listener's gazing ratios we couldn't find a significant interaction between language difference and expertise in L2.

To investigate the issue further, we conducted ANOVA tests to evaluate the effect of the volume of data on analysis of the interaction between language difference and expertise in L2, adding incrementally data of every two sets to half of the data (10 sets). Figure 11 depicts significance probabilities of interaction between language difference and expertise in L2 that were calculated using both the expertise of the speaker (Gaze_at_Rank) and the eye gaze directions from participants of higher expertise to participants of lower expertise and those of participants of lower expertise to participants of higher expertise (Gaze_of_H2L, L2H). The datasets were numbered according to the order of data collection. As

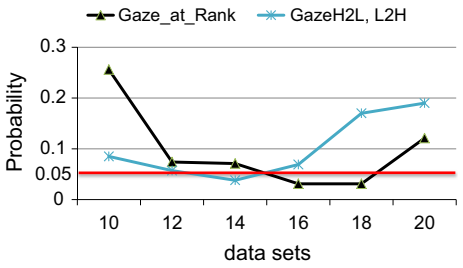


**Fig. 11** Probabilities of interactions between language difference and expertise in L2 for listener's gazing ratios at each Rank (Gaze_at_Rank) and those from high to low and low to high (GazeH2L, H2L)

shown in Fig. 11, significance probabilities of the interaction became lower as the volume of data increased due to the increased number of sets and were under the probability 0.05 (a significant interaction level) in both cases. The significance probabilities, however, again increased in both cases as more datasets were used.

These results suggested that the volume of the multimodal corpus was still insufficient to analyze interaction in listener's gazing activities between language difference and expertise in L2, thus implying the possibility that there might be interaction between language difference and expertise in L2 (refer to Appendix 2).

We analyzed features of utterances and eye gaze of the collected multimodal data only from the stochastic viewpoint in this paper but did not conduct any detailed analysis of the relation between eye gaze and utterances as done for conversations in L1 by Goodwin (1979). Our transcription showed that the participants tended to produce not only simpler expressions but also imperfect or fragmental ones more often in English than in Japanese conversations, probably because of their low language proficiency. We plan to analyze the differences in disfluencies between L1 and L2 conversations and their relation with eye gaze behaviors, although it requires an additional major annotation effort.

## 6 Conclusion

We collected an 8-hour multimodal corpus of multiparty conversations to investigate the differences in communicative activities by the same interlocutors in Japanese (their L1) and in English (their L2). Although annotation for speech acts and alignment of transcribed speech data are still being conducted, we found some interesting features of utterance and eye gaze by analyzing the conversational data in which annotations on speech, gaze events, and turn-taking activity were completed.

We confirmed that the total and average utterance durations were shorter and that the silence in the L2 conversations was longer than those in the L1 conversations, which suggested the difficulties experienced by the participants in conversations in L2. We also confirmed that eye gazes in the L2 conversations were different from

those in L1. Speakers were gazed at more by listeners in conversations in L2 than in L1. The reason why the speaker was gazed at more by listeners in conversations in L2 than those in L1 is still not clear, and analyses of the relations between the gaze activities and speech acts in speaking would seem necessary to clarify this point.

The results on the total utterance durations and the average utterance durations among the participants revealed significant interactions between the expertise in L2 and the language difference. As for eye gaze, we could not find a significant interaction between language difference and expertise in L2, but the experimental results using subsets of the data suggested the possibility that there might be interaction between language difference and expertise in L2.

Considering that second-language conversations are commonly observed in daily life throughout the world, our findings on differences in conversations in L1 and L2 suggest possible directions for future research in psychology and cognitive science as well as in human–computer interaction technologies. This study provided a basis for monitoring the status of all participants by studying the effects of their linguistic proficiency on communicative activities and attitudes in second-language conversations. These results can be used to support mutual understanding and balanced participant contributions, in cases of uneven linguistic proficiencies, in cooperative activities involving computer-supported cooperative work (CSCW). The results may also be important in developing humanoid robots or agents for dialogue-based computer-assisted language learning.

We plan to collect more multimodal data and continue annotation and transcription of multimodal corpora to obtain more reliable results on the differences between participants of different expertise in L2. Moreover, we intend to make the multimodal corpora available to the research community after completing the annotation work.

**Acknowledgments** The authors would like to thank Professor Kristiina Jokinen of the University of Helsinki and Emeritus Professor Masuzo Yanagida of Doshisha University for their suggestions and the various discussions we had with them. The authors would also like to thank Kosuke Kabashima, Shota Yamasaki, and Satoko Nomoto of Doshisha University for their efforts in collecting multimodal conversational data. This research was supported by a contract with MEXT number 245000336.

# Appendix 1: Analyses of second order interactions

==========

Total Utterance Duration:

==========

2nd Order Interaction:
- LD x CD x ED: $F_{(2, 57)} = 5.01$, $p = .010$

---

- Simple Interaction: LD x ED
  GO: $F_{(2, 57)} = 10.23$, $p = .000$
- Simple Interaction: CD x ED
  ENG: $F_{(2, 57)} = 4.30$, $p = .018$
  JPN: $F_{(2, 57)} = 3.63$, $p = .033$
- Simple Interaction: LD x CD
  Rank3: $F_{(1, 57)} = 6.27$, $p = .015$

---

- Simple Simple Main Effect: ED
  ENG x FF: $F_{(2, 57)} = 2.53$, $p = .088$
  ENG x GO: $F_{(2, 57)} = 6.22$, $p = .004$
- Simple Simple Main Effect: LD
  Rank1 x FF: $F_{(1, 57)} = 26.29$, $p = .000$
  Rank2 x FF: $F_{(1, 57)} = 19.54$, $p = .000$
  Rank3 x FF: $F_{(1, 57)} = 51.27$, $p = .000$
  Rank1 x GO: $F_{(1, 57)} = 8.02$, $p = .006$
  Rank2 x GO: $F_{(1, 57)} = 3.99$, $p = .050$
  Rank3 x GO: $F_{(1, 57)} = 62.51$, $p = .000$
- Simple Simple Main Effect: CD
  Rank2 x ENG: $F_{(1, 57)} = 4.14$, $p = .046$
  Rank3 x ENG: $F_{(1, 57)} = 11.30$, $p = .001$
  Rank2 x JPN: $F_{(1, 57)} = 11.80$, $p = .001$

---

Multiple Comparison Test: ED
(with Bonferroni correction: p = .0167)
- ENG x FF:
  Rank2 vs. Rank1: not significant
  Rank3 vs. Rank1: not significant
  Rank2 vs. Rank3: $F_{(1, 57)} = 4.81$, $p = .032$ (marginally significant)
- ENG x GO:
  Rank2 vs. Rank1: not significant
  Rank3 vs. Rank1: $F_{(1, 57)} = 11.06$, $p = .002$
  Rank2 vs. Rank3: $F_{(1, 57)} = 7.18$, $p = .010$
- JPN x FF:
  Rank2 vs. Rank1: not significant
  Rank3 vs. Rank1: not significant
  Rank2 vs. Rank3: not significant
- JPN x GO:
  Rank2 vs. Rank1: not significant
  Rank3 vs. Rank1: not significant
  Rank2 vs. Rank3: not significant

==========

Average Utterance Duration:

==========

2nd Order Interaction:
- Language x CD x ED: $F_{(2, 57)} = 4.450$, $p = .016$

---

- Simple Interaction: LD x ED
  GO: $F_{(2, 57)} = 4.96$, $p = .010$
- Simple Interaction: CD x ED
  ENG: $F_{(2, 57)} = 2.84$, $p = .067$
- Simple Interaction: LD x CD
  Rank3: $F_{(1, 57)} = 6.70$, $p = .012$

---

- Simple Simple Main Effect: LD
  Rank1 x FF: $F_{(1, 57)} = 10.71$, $p = .002$
  Rank2 x FF: $F_{(1, 57)} = 5.31$, $p = .025$
  Rank3 x FF: $F_{(1, 57)} = 6.32$, $p = .015$
  Rank1 x GO: $F_{(1, 57)} = 3.95$, $p = .052$
  Rank3 x GO: $F_{(1, 57)} = 26.33$, $p = .000$

## Appendix 2: Calculation of average gazing ratio

There are two methods of calculating the averages of the gazing ratios. For example, the average of the speaker's gazing ratios can be calculated in two ways:

$$\textit{Average of speaker's gazing ratios} = \frac{1}{n}\sum_{i=1}^{n}\frac{DSG_j(i)}{D(i)} \times 100\,(\%),$$

and

$$\textit{Average of speaker's gazing ratios} = \frac{\sum_{i=1}^{n} DSG_j(i)}{\sum_{i=1}^{n} D(i)} \times 100\,(\%).$$

The former and the latter are referred to as the macro-average and the micro-average, respectively, in the field of information retrieval.

In this paper we used the macro-average method to calculate the averages of both the speaker's and the listener's gazing ratios, assuming that each utterance is equally important for participants because we did not focus on the meaning or function of each utterance from the perspective of discourse in this paper. The figures calculated based on the micro-average are listed in Tables 7 and 8, which correspond to the figures in Tables 5 and 6, respectively.

To Provide a detailed explanation of the averaging over each conversation group, the averages of the speaker's and listener's gazing ratios are given by

$$\textit{Average of speaker's gazing ratios} = \frac{1}{6}\left(\sum_{j=1}^{3}\sum_{\substack{k=1\\k\neq j}}^{3} SGR(j,k)\right)$$

**Table 7** Basic statistics of eye gaze activities

| Features in conversation | Average ± SD | | | |
|---|---|---|---|---|
| | Free (JPN) | Free (ENG) | Goal (JPN) | Goal (ENG) |
| Speaker's gazing ratios | 0.28 ± 0.13 | 0.27 ± 0.14 | 0.28 ± 0.16 | 0.27 ± 0.17 |
| Listener's gazing ratios (LGR) | 0.55 ± 0.15 | 0.64 ± 0.15 | 0.50 ± 0.16 | 0.63 ± 0.16 |
| LGR at Rank1 | 0.54 ± 0.15 | 0.62 ± 0.12 | 0.51 ± 0.16 | 0.60 ± 0.17 |
| LGR at Rank2 | 0.53 ± 0.17 | 0.66 ± 0.14 | 0.49 ± 0.14 | 0.66 ± 0.15 |
| LGR at Rank3 | 0.57 ± 0.12 | 0.63 ± 0.19 | 0.51 ± 0.18 | 0.62 ± 0.17 |
| LGR from HIGH to LOW | 0.54 ± 0.16 | 0.63 ± 0.13 | 0.50 ± 0.16 | 0.61 ± 0.16 |
| LGR from LOW to HIGH | 0.57 ± 0.13 | 0.64 ± 0.17 | 0.50 ± 0.16 | 0.64 ± 0.16 |

**Table 8** Features of ANOVA test results on eye gaze activities

| Features | ANOVA test results | | |
|---|---|---|---|
| | Language diff. (LD) | Conversation type diff. (CD) | LD, CD, and diff. of expertise in L2 (ED) |
| Average of speaker's gazing ratios | n.s.m. ($F_{(1,117)} = 0.65$) | n.s.m ($F_{(1,117)} = 0.17$) | n.s.i. ($F_{(2,117)} = 1.4$) between LD and ED |
| Average of listener's gazing ratios (LGR at Rank) | $F_{(1,117)} = 84.8$ $P < 0.01$ | $F_{(1,117)} = 7.4$, $P < 0.01$ | $F_{(2,117)} = 4.2$, $P < 0.05$ between LD and ED |
| Average of listener's gazing ratios (LGR from H to L/from L to H) | $F_{(1,117)} = 81.6$, $P < 0.01$ | $F_{(1,117)} = 7.6$, $P < 0.01$ | n.s.i. ($F_{(2,117)} = 0.18$) between LD and ED |

$$\textit{Average of listener's gazing ratios} = \frac{1}{6}\left(\sum_{j=1}^{3}\sum_{\substack{k=1\\k\neq j}}^{3} LGR(j,k)\right)$$

Here, $SGR(j, k)$ indicates the average of the speaker's gazing ratios in the case where the $j$th participant is the speaker and the $k$-th participant is a listener, and $LGR(j, k)$ shows the average of the listener's gazing ratios in the case where the $j$th participant is a listener and the $k$th participant is the speaker.

The tendencies of the basic statistics and the features of the ANOVA results are almost the same as those obtained based on the macro-averages shown in Tables 5 and 6, except for the ANOVA result of the listener's gazing ratios (LGR at Rank). These results suggest that we may not be able to regard each utterance as equally important in calculating the average value of eye gaze activities, thus implying the possibility that there might be interaction between language difference and expertise in L2.

## References

Allwood, J., Cerrato, L., Jokinen, K., Navarreta, C., & Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management, and sequencing phenomena. *International Journal of Language Resources and Evaluation, 41*(4), 273–287.

Anguera, X., Bozonnet, S., Evans, N., Fredoulle, C., & Friedland, G. (2012). Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech and Language Processing, 20*(2), 356–370.

Argyle, M., & Cook, M. (1976). Gaze and mutual gaze. Cambridge: Cambridge University Press.

Argyle, M., Lalljee, M., & Cook, M. (1968). The effects of visibility on interaction in dyad. *Human Relations, 21*, 3–17.

Beattie, G. W. (1978). Floor apportionment and gaze in conversational dyads. *British Journal of Social and Clinical Psychology, 17*, 7–16.

Beattie, G. W. (1980). The role of language production process in the organization of behaviour in face-to-face interaction. In B. Butterworth (Ed.), *Language production* (Vol. 1, pp. 69–107). London: Academic Press.

Beyene, T., Pamela, J., Hinds, P. J., & Cramton, C. D. (2009). Walking through jelly: language proficiency, emotions, and disrupted collaboration in global work. *SSRN eLibrary*.

Boyle, E., Anderson, A., & Newlands, A. (1994). The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language and Speech*, *37*, 1–20.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemont, M., Hain, T., Kaldec, J., Karaiskov, V., Kraaji, W., Kronenthat, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., & Wellner, P. (2005). The AMI meeting corpus: A pre-announcement. In *Workshop on machine learning for multimodal interaction (MLMI'05)*, Edinburgh, UK.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*, 62–81.

ELAN. http://tla.mpi.nl/tools/tla-tools/elan/

ELAN. http://www.lat-mpi.eu/tools/elan

Garofolo, J., Laprun, C., Michel, M., Stanford, V., & Tabassi, E. (2004). The NIST meeting room pilot corpus. In *International conference on language resources and evaluation (LREC2004)*, Lisbon, Portugal.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *TRENDS in Cognitive Sciences, 8*, 8–11.

Gatica-Perez, D. (2009). Automatic nonverbal analysis of social interaction in small groups: a review. *Image and Vision Computing, Special Issue on Human Behavior, 27*(12), 1775–1787.

Goffman, E. (1976). Replies and responses. *Language in Society, 5*, 257–313.

Goodwin, C. (1979). The interactive construction of a sentence in natural conversation. In G. Psathas (Ed.), *Everyday language: studies in ethnomethodology* (pp. 97–121). New York: Irvington Publishers.

Hain, T., Burget, L., Dines, J., Garner, P. N., Hannani, A. E., Hujibregts, M., et al. (2010). The AMIDA 2009 meeting transcription system. *Proceedings of INTERSPEECH, 2010*, 358–361.

Hosoda, Y. (2006). Repair and relevance of differential language expertise in second language conversations. *Applied Linguistics, 27*, 25–50.

Jokinen, K., Furukawa, H., Nishida, M., & Yamamoto, S. (2013). Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *3*(2), 1–30.

Kabashima, K, Nishida, M., Jokinen, K., & Yamamoto, S. (2012). Multimodal corpus of conversations in mother tongue and second language by same interlocutors. In *Proceedings of 4th workshop on eye gaze in intelligent human machine interaction*, Santa Fe, USA.

Kasper, G. (2004). Participant orientations in conversations-for-learning. *The Modern Language Journal, 88*, 551–567.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica, 26*, 22–63.

Kleinke, C. L. (1986). Gaze and eye contact: a research review. *Psychological Bulletin, 100*, 78–100.

Mehrabian, A., & Ferris, S. R. (1967). Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology, 31*(3), 248–252.

Mehrabian, A., & Wiener, M. (1967). Decoding of inconsistent communications. *Journal of Personality and Social Psychology, 6*(1), 109–114.

Moattar, M. H., & Homayounpour, M. M. (2012). A review on speaker diarization systems and approaches. *Speech Communication, 54*, 1065–1103.

MPIP. http://www.mpi.nl/

NAC Image Technology Inc. http://www.nacinc.jp/

Pentland, A. (2005). Socially aware computation and communication. *IEEE Computer, 38*(3), 33–40.

TOEIC. http://www.ets.org/toeic

TOEIC_TEST. http://www.toeic.or.jp/toeic/about/data/data_avelist.html

Tur, G., Stolcke, A., Voss, L., Dowing, J., Favre, B., Fernandez, R., et al. (2008). The CALO meeting speech recognition and understanding system. *Proceedings of Spoken Language Technology Workshop, 2008*, 69–72.

Umata, I., Yamamoto, S., Ijuin, K., & Nishida, M. (2013). Effects of language proficiency on eye-gaze in second language conversations: Toward supporting second language collaboration. In *The international conference on multimodal interaction ICMI2013*, pp. 413–419. Sydney, Australia.

Vinciarelli, A., Patic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing, 27*(12), 1743–1759.

Yamamoto, S., Taguchi, K., Umata, I., Kabashima, K., & Nishida, M. (2013). Differences in interactional attitudes in native and second language conversations: Quantitative analyses of multimodal three-party corpus. In *Proceedings of 35th annual conference of the cognitive science (CogSci2013)*. Berlin, Germany.

Yamasaki, S., Furukawa, H., Nishida, M., Jokinen, K., & Yamamoto, S. (2012). Multimodal corpus of multi-party conversations in second language. In: *International conference on language resources and evaluation (LREC2012)*, Istanbul, Turkey.